Long-form factuality in large language models

Year: 2024 | Citations: 101 | Authors: Jerry Wei, Chengrun Yang, Xinying Song

---

Abstract

Large language models (LLMs) often generate content that contains factual errors when responding to fact-seeking prompts on open-ended topics. To benchmark a model's long-form factuality in open domains, we first use GPT-4 to generate LongFact, a prompt set comprising thousands of questions spanning 38 topics. We then propose that LLM agents can be used as automated evaluators for long-form factuality through a method which we call Search-Augmented Factuality Evaluator (SAFE). SAFE utilizes an LLM to break down a long-form response into a set of individual facts and to evaluate the accuracy of each fact using a multi-step reasoning process comprising sending search queries to Google Search and determining whether a fact is supported by the search results. Furthermore, we propose extending F1 score as an aggregated metric for long-form factuality. To do so, we balance the percentage of supported facts in a response (precision) with the percentage of provided facts relative to a hyperparameter representing a user's preferred response length (recall). Empirically, we demonstrate that LLM agents can outperform crowdsourced human annotators - on a set of ~16k individual facts, SAFE agrees with crowdsourced human annotators 72% of the time, and on a random subset of 100 disagreement cases, SAFE wins 76% of the time. At the same time, SAFE is more than 20 times cheaper than human annotators. We also benchmark thirteen language models on LongFact across four model families (Gemini, GPT, Claude, and PaLM-2), finding that larger language models generally achieve better long-form factuality. LongFact, SAFE, and all experimental code are available at https://github.com/google-deepmind/long-form-factuality.